WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Resume and CV Screening Using NLP

Mallela Chandrakanth[1], Shri N. Naveen Kumar[2]

[1](Post Graduate Student, M. Tech (SE) Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad,
Email: chandumallela007@gmail.com)
[2](Associate Professor, Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad,
Email: naveen.cse.mtech@gmail.com)

ABSTRACT

In today's digital age, organizations often struggle to efficiently manage the overwhelming number of job applications they receive. Manual resume screening processes are not only labor-intensive and slow but also prone to human error and bias. This study introduces an AI-driven resume screening system that utilizes Natural Language Processing (NLP) and Machine Learning (ML) to streamline and improve candidate assessment. By extracting and analyzing data from resumes and comparing it with job descriptions, the system generates relevance scores and ranks candidates accordingly. The proposed solution delivers fast, accurate, and objective results, making it highly effective for modern recruitment needs.

Keywords — AI in Recruitment, NLP, Resume Analysis, Machine Learning, Candidate Scoring, Hiring Automation

INTRODUCTION

The recruitment process is a critical component of every organization's success, and selecting the right candidate from a large pool of applicants is often challenging. Traditional methods of resume screening rely heavily on human effort, making them slow, inconsistent, and often subject to unconscious bias. As companies continue to receive hundreds or even thousands of applications for a single role, the demand for an efficient, accurate, and scalable solution has become more pressing than ever.

With the advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP), it is now possible to automate various aspects of candidate evaluation. In this project, we propose a smart resume screening system that leverages NLP to extract meaningful information from resumes and job descriptions. By applying topic modeling and similarity scoring techniques such as Latent Dirichlet Allocation (LDA) and TF-IDF with cosine similarity, the system identifies key topics in the job description and evaluates how well each resume aligns with those requirements.

The system not only improves the efficiency of the screening process but also reduces human bias and enhances the overall quality of candidate shortlisting. By ranking resumes based on relevance to the job role, it provides recruiters with a data-driven approach to decision-making. This AI-based solution can be integrated into existing hiring workflows to support faster, fairer, and more consistent talent acquisition in a wide range of industries.

LITERATURE REVIEW

In recent years, the increasing volume of digital job applications has motivated researchers to explore automated approaches for resume screening. Traditional recruitment processes involve manually reviewing resumes, which is time-consuming and often leads to inconsistencies in candidate evaluation. Studies have shown that automation not only speeds up the hiring process but also improves the quality of candidate filtering by focusing on relevant skills and qualifications rather than subjective factors.

Several works have explored the use of Natural Language Processing (NLP) techniques to extract information from unstructured resume data. Techniques such as tokenization, stemming, and stop word removal are commonly used to process and clean textual data. Researchers have employed Named Entity Recognition (NER) and keyword extraction to identify critical elements like skills, education, and experience. This information is then used to match resumes with job descriptions in a structured and meaningful way.

Machine Learning models have also played a significant role in resume analysis. Classification algorithms such as Support Vector Machines (SVM), Decision Trees, and Random Forests have been used to categorize resumes based on job fit. More recently, topic modeling techniques like Latent Dirichlet Allocation (LDA) have been adopted to discover hidden themes in text, which helps in understanding the core content of resumes and job descriptions. These models provide insights into the alignment of candidate profiles with job expectations.

Cosine similarity, along with TF-IDF (Term Frequency–Inverse Document Frequency), is widely used in information retrieval systems and has been applied effectively in resume-job matching tasks. These techniques assess the relevance of a resume by comparing it to a job description in terms of textual similarity. Researchers have found that combining statistical text representation with semantic understanding leads to more accurate candidate rankings.

Despite the promising advancements, challenges such as domain-specific language, varied resume formats, and contextual differences still exist. However, the integration of AI-powered tools into recruitment systems continues to grow. Literature emphasizes that when NLP and ML techniques are thoughtfully applied, they can significantly reduce bias, increase recruitment efficiency, and improve the overall quality of hires. This forms a strong foundation for developing practical and scalable AI-based resume screening systems.

EXISTING SYSTEM

Several recruitment platforms and HR software tools already incorporate automation and artificial intelligence to enhance resume screening. For instance, Applicant Tracking Systems (ATS) like **Workday, Taleo, and Greenhouse** are widely used by organizations to manage and sort applications. These platforms typically rely on keyword-based filters to match resumes with job descriptions. While effective for basic screening, these systems often fail to understand the context or deeper semantics of a candidate's experience, leading to false negatives — qualified candidates being overlooked due to missing keywords or unconventional phrasing.

Advanced tools such as **HireVue, Pymetrics, and SeekOut** attempt to integrate machine learning and data-driven insights into recruitment. Some utilize NLP to extract candidate information, while others incorporate video analysis or psychometric assessments. However, these systems are often expensive, proprietary, and may not provide full transparency on their decision-making processes. Furthermore, many existing solutions focus more on enterprise-level users and lack customizable features tailored for specific organizational needs or niche job profiles. This opens a gap for lightweight, open-source, and transparent AI solutions that can be adapted and scaled for diverse hiring needs.

PROPOSED SYSTEM

The proposed system introduces a modern and intelligent approach to resume screening by integrating **Natural Language Processing (NLP)** and **Machine Learning (ML)** to evaluate job applicants. Unlike traditional keyword-based filters, this system analyzes the actual content and structure of resumes to understand the context and relevance of the applicant's skills, experience, and qualifications in relation to the job description. This is achieved through a combination of **TF-IDF vectorization** and **Latent Dirichlet Allocation (LDA)** topic modeling to ensure that deeper textual patterns and semantic matches are identified.

A core feature of the system is its **resume parsing and preprocessing pipeline**. Uploaded resumes in PDF format are first converted into plain text using tools like pdfplumber. The text is then cleaned and tokenized, removing stop words and irrelevant characters. Preprocessed data is passed through topic modeling algorithms like LDA, which extracts key topics from resumes and compares them with topics from job descriptions. This ensures that even resumes with non-standard formats or phrasing can be accurately evaluated.

In addition to topic modeling, the system employs **TF-IDF and cosine similarity** to compute how closely a resume matches the job description. This dual approach – combining both content similarity and topic relevance – enables a more balanced evaluation of resumes. Candidates are ranked based on their similarity scores, and the system highlights missing but relevant keywords from the job description that are not found in the resume. This information can be valuable for both recruiters and candidates, providing transparency in how decisions are made.

The system is built as a **user-friendly Streamlit web application**, making it easy for HR professionals to upload resumes and input job descriptions without any technical expertise. For each resume, the tool displays extracted topics, missing topic words, and a similarity score in a clear and organized manner. Resumes are also sorted automatically based on their relevance scores, allowing recruiters to focus on the most suitable candidates first, reducing the time and effort required for manual screening.

Overall, the proposed system aims to enhance fairness, efficiency, and accuracy in the recruitment process. By automating the analysis of resumes with intelligent models, it minimizes human bias and ensures that candidates are evaluated based on their actual capabilities and alignment with job requirements. This system is also scalable, cost-effective, and adaptable to various domains, making it a practical solution for organizations seeking to modernize their hiring workflows.

ARCHITECTURE

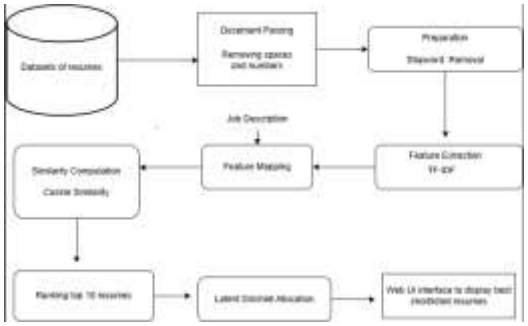

Figure No 1: Architecture

MODEL

The core model of the resume screening system combines **Natural Language Processing (NLP)** and **Machine Learning (ML)** techniques to accurately evaluate the relevance of resumes against job descriptions. The model functions in multiple stages, starting with resume parsing, where uploaded PDF documents are converted to plain text using a PDF extraction library. This raw text is then cleaned, tokenized, and filtered to remove stop words and non-alphabetic terms. This step ensures that only meaningful words are retained for further analysis.

Once the text is preprocessed, two distinct models are applied: **TF-IDF (Term Frequency-Inverse Document Frequency)** and **LDA (Latent Dirichlet Allocation)**. The TF-IDF model transforms text data into numerical vectors that reflect the importance of each word in a resume relative to the job description. These vectors are then used to compute **cosine similarity**, which measures the degree of textual overlap between resumes and the job description. This provides a score that indicates how closely each resume aligns with the employer's requirements.

In parallel, the LDA model performs **topic modeling** to uncover hidden themes or topics present in the resumes and job description. This adds a layer of semantic analysis by identifying word clusters that represent broader skill areas or domains. By comparing the extracted topics from resumes with those in the job description, the system highlights missing but relevant terms, helping recruiters understand where a resume might fall short even if keywords partially match.

The final model output consists of a **ranked list of resumes**, each annotated with a similarity score, extracted topics, and a list of missing topic keywords. This multi-layered analysis ensures that resumes are not only matched based on surface-level terms but also evaluated for their deeper relevance and thematic alignment. The combination of TF-IDF and LDA allows the system to deliver both precision and context-awareness in resume evaluation, offering recruiters a more informed and efficient screening process.

ALGORITHM

The resume screening system relies on a **hybrid algorithm** that integrates two powerful techniques: **TF-IDF with cosine similarity** for direct keyword matching and **Latent Dirichlet Allocation (LDA)** for topic-based semantic analysis. The first step of the algorithm involves **text extraction and preprocessing**, where PDF resumes and job descriptions are converted into text. This text is then cleaned by removing stop words, converting to lowercase, and filtering out non-alphabetic characters to ensure consistency and relevance in the processed data.

Next, the algorithm applies **TF-IDF vectorization** to both the job description and each resume. This technique calculates the importance of words in a document relative to a set of documents, highlighting terms that are significant in a specific resume but not common across all. Using **cosine similarity**, the algorithm then compares the job description vector with each resume vector. The result is a numeric score that indicates how similar each resume is to the job description based on shared keywords and their importance.

In parallel, the algorithm runs **LDA topic modeling** to identify dominant topics within the job description and resumes. This helps capture the **semantic meaning** and context behind the text, rather than just exact word matches. Each resume is evaluated based on how well its topics align with those of the job description, and missing topic terms are identified. Combining the outputs of cosine similarity and topic relevance, the algorithm ranks the resumes and presents detailed feedback on strengths and gaps. This ensures both **surface-level accuracy and deeper contextual matching** during the resume screening process.

[1] Five Steps of Automated Resume Screening:

1. **Resume and Job Description Input**

   The system begins by collecting resumes in PDF format and a relevant job description from the recruiter. Resumes can be uploaded through the user interface, and the job description is entered in text form. These documents serve as the core input for the screening process.

2. **Text_Preprocessing**

   The uploaded documents are then preprocessed to enhance the quality of text analysis. This step includes converting text to lowercase, removing punctuation and stop words, and filtering out non-alphabetic characters. Tokenization is used to break down text into individual words, enabling accurate comparison in later stages.

3. **Feature Extraction and Topic Modeling**

   Each document undergoes two kinds of feature extraction. First, TF-IDF is applied to capture term relevance across documents. Second, Latent Dirichlet Allocation (LDA) is used to uncover hidden topics within the resumes and the job description. These topics help identify thematic similarities between the job requirements and candidates' skills.

4. **Similarity Computation and Ranking**

   Cosine similarity is calculated between the TF-IDF vector of the job description and each resume. A numerical score is generated that quantifies how closely each resume matches the job requirements. Simultaneously, LDA topic overlap is analyzed to highlight missing keywords in each resume. Resumes are then ranked based on both similarity and topic coverage.

5. **Filtering and Shortlisting**

   Candidates whose resumes score below a certain threshold can be flagged for low relevance. This allows recruiters to focus on the top-ranked candidates whose resumes align well with the job description both in terms of keywords and thematic content. The system provides detailed insights, such as missing topic terms, to assist in the final decision-making process.

RESULTS

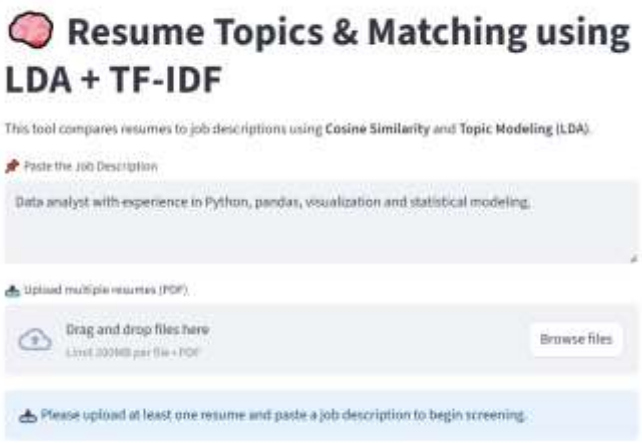

Figure No 2: Before uploading the resumes

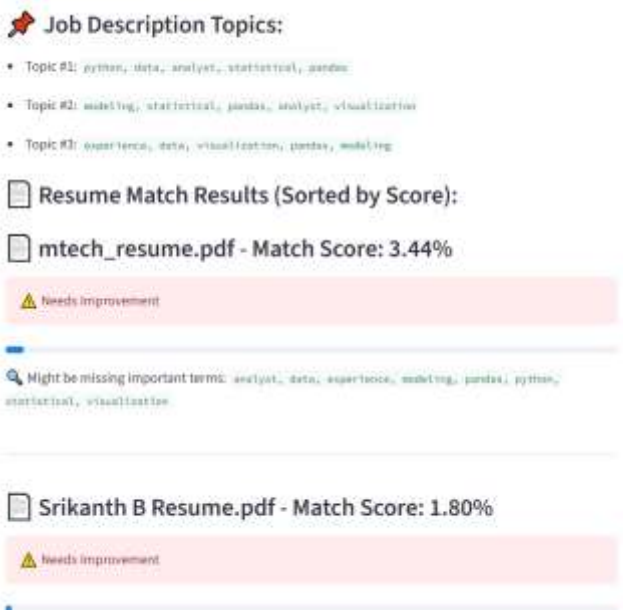

Figure No 3: After uploading the resumes

CONCLUSION

In today's competitive job market, manual resume screening presents a major bottleneck for recruiters and HR professionals. It is not only time-consuming but also vulnerable to subjective judgment and human errors. To address these limitations, this project introduces an automated resume screening system that leverages Natural Language Processing (NLP) and Machine Learning techniques to streamline the candidate evaluation process. By automating repetitive screening tasks, the system helps reduce recruiter workload and speeds up the hiring pipeline.

The proposed system integrates TF-IDF-based similarity analysis and Latent Dirichlet Allocation (LDA) for topic modeling. This hybrid approach enables the system to analyze not just keyword matches but also the underlying topics and themes in resumes and job descriptions. It ensures that applicants are evaluated based on content relevance and semantic alignment rather than simple keyword counts. As a result, candidates with the most contextually appropriate skills and qualifications are ranked higher.

One of the key strengths of the system lies in its ability to identify missing but crucial topic terms in a candidate's resume. These insights are highly valuable for both recruiters and applicants. Recruiters gain visibility into potential skill gaps, while job seekers can better tailor their resumes to meet role-specific expectations. This level of interpretability sets the system apart from traditional black-box models.

The project also emphasizes user-friendliness and practical applicability. The interface built using Streamlit allows users to upload multiple resumes and view similarity scores, topic matches, and keyword gaps in a structured format. This intuitive design ensures that even non-technical recruiters can effectively use the tool without prior training. The modular architecture also allows future upgrades, such as integrating named entity recognition or support for multilingual resumes.

In summary, the AI-driven resume screening system developed in this project demonstrates how intelligent automation can transform recruitment practices. It offers a fair, scalable, and efficient solution to modern hiring challenges. By combining linguistic analysis and machine learning, the tool provides deeper candidate insights and enhances decision-making for recruiters. With further refinements, such systems could become standard tools in HR departments globally.

REFERENCES

[1]. Bhatia, S. (2015). *Resume Parsing and Matching using NLP Techniques*. International Journal of Computer Applications, 116(11), 1–5.

[2]. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781.

[3]. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3(Jan), 993–1022.

[4]. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

[5]. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

[6]. Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv preprint arXiv:1908.10084.

[7]. Gensim. (n.d.). *Topic Modeling with LDA*. Retrieved from https://radimrehurek.com/gensim/

[8]. Streamlit Inc. (n.d.). *Streamlit Documentation*. Retrieved from https://docs.streamlit.io/

[9]. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). *Attention is All You Need*. Advances in Neural Information Processing Systems, 30.

[10]. Chawla, D., & Bansal, S. (2020). *Automated Resume Screening Using Machine Learning*. International Journal of Computer Applications, 176(30), 15–19.